Title

Exploring the loblolly pine (Pinus taeda L.) genome by BAC sequencing and Cot analysis

Authors

Perera, Dinum
Magbanua, Zenaida V
Thummasuwan, Supaphan
et al.

Publication Date

2018-07-01

DOI

10.1016/j.gene.2018.04.024

Peer reviewed

Gene 663 (2018) 165–177

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Exploring the loblolly pine (*Pinus taeda* L.) genome by BAC sequencing and Cot analysis

Dinum Perera[a], Zenaida V. Magbanua[b], Supaphan Thummasuwan[c], Dipaloke Mukherjee[d], Mark Arick II[a], Philippe Chouvarine[e], Campbell J. Nairn[f], Jeremy Schmutz[h,i], Jane Grimwood[h,i], Jeffrey F.D. Dean[g], Daniel G. Peterson[a,j,*]

[a] Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, MS 39762, USA
[b] National Institute of Molecular Biology & Biotechnology, National Science Complex, College of Science, University of the Philippines, Diliman, Quezon City, Philippines
[c] Department of Agricultural Sciences, Naresuan University, Phitsanulok, Thailand
[d] Department of Food Science, Nutrition, & Health Promotion, Mississippi State University, Mississippi State, MS 39762, USA
[e] Texas Children's Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA
[f] Warnell School of Forest Resources, University of Georgia, Athens, GA 30602, USA
[g] Department of Biochemistry, Molecular Biology, Entomology & Plant Pathology, Mississippi State University, Mississippi State, MS 39762, USA
[h] US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA
[i] HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, AL 35801, USA
[j] Department of Plant & Soil Sciences, Mississippi State University, Mississippi State, MS 39762, USA

ARTICLE INFO

Keywords:
Pinus taeda
Loblolly pine
Bacterial artificial chromosome
BAC
Cot analysis
Carbon metabolism

ABSTRACT

Loblolly pine (LP; *Pinus taeda* L.) is an economically and ecologically important tree in the southeastern U.S. To advance understanding of the loblolly pine (LP; *Pinus taeda* L.) genome, we sequenced and analyzed 100 BAC clones and performed a Cot analysis. The Cot analysis indicates that the genome is composed of 57, 24, and 10% highly-repetitive, moderately-repetitive, and single/low-copy sequences, respectively (the remaining 9% of the genome is a combination of fold back and damaged DNA). Although single/low-copy DNA only accounts for 10% of the LP genome, the amount of single/low-copy DNA in LP is still 14 times the size of the *Arabidopsis* genome. Since gene numbers in LP are similar to those in *Arabidopsis*, much of the single/low-copy DNA of LP would appear to be composed of DNA that is both gene- and repeat-poor. Macroarrays prepared from a LP bacterial artificial chromosome (BAC) library were hybridized with probes designed from cell wall synthesis/wood development cDNAs, and 50 of the "targeted" clones were selected for further analysis. An additional 25 clones were selected because they contained few repeats, while 25 more clones were selected at random. The 100 BAC clones were Sanger sequenced and assembled. Of the targeted BACs, 80% contained all or part of the cDNA used to target them. One targeted BAC was found to contain fungal DNA and was eliminated from further analysis. Combinations of similarity-based and ab initio gene prediction approaches were utilized to identify and characterize potential coding regions in the 99 BACs containing LP DNA. From this analysis, we identified 154 gene models (GMs) representing both putative protein-coding genes and likely pseudogenes. Ten of the GMs (all of which were specifically targeted) had enough support to be classified as intact genes. Interestingly, the 154 GMs had statistically indistinguishable ($\alpha = 0.05$) distributions in the targeted and random BAC clones (15.18 and 12.61 GM/Mb, respectively), whereas the low-repeat BACs contained significantly fewer GMs (7.08 GM/Mb). However, when GM length was considered, the targeted BACs had a significantly greater percentage of their length in GMs (3.26%) when compared to random (1.63%) and low-repeat (0.62%) BACs. The results of our study provide insight into LP evolution and inform ongoing efforts to produce a reference genome sequence for

Abbreviations: LP, loblolly pine; GMs, gene models; BAC, bacterial artificial chromosome; SPB, sodium phosphate buffer; Tm, melting temperature; M_{mvc}, monovalent cation concentration; M·s, product of nucleotide molarity and reassociation time in seconds; HR, highly repetitive; MR, moderately repetitive; SL, single/low-copy; ORF, open reading frame; LTR, long terminal repeat; GMP, GDP-D-mannose pyrophosphorylase; CCoAOMT, caffeoyl-CoA *O*-methyltransferase; LAC, laccase; KOR, Korrigan endoglucanase; CesA, cellulose synthase; PAL, phenylalanine ammonia lyase; SuSy, sucrose synthase; MYB, MYB transcription factor; AED, annotation edit distance; Csl, cellulose synthase-like; H, *p*-hydroxyphenyl; G, guaiacyl; S, syringyl; HAP, hydroxyapatite

* Corresponding author at: Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, MS 39762, USA.
E-mail addresses: supaphant@nu.ac.th (S. Thummasuwan), dm259@msstate.edu (D. Mukherjee), maa146@igbb.msstate.edu (M. Arick), nairn@uga.edu (C.J. Nairn), jschmutz@hudsonalpha.org (J. Schmutz), jgrimwood@hudsonalpha.org (J. Grimwood), jeffdean@bch.msstate.edu (J.F.D. Dean), dp127@msstate.edu (D.G. Peterson).

https://doi.org/10.1016/j.gene.2018.04.024
Received 5 October 2017; Received in revised form 20 March 2018; Accepted 10 April 2018
Available online 12 April 2018

LP, while characterization of genes involved in cell wall production highlights carbon metabolism pathways that can be leveraged for increasing wood production.

1. Introduction

Loblolly pine (LP; *Pinus taeda* L.; $2n = 24$), is a highly important commercial tree species in the United States (Plomion et al., 2007). It is the leading forest tree for timber, pulp, and paper industries in the southeastern U.S. (Plomion et al., 2007), and because of its relatively rapid growth, there is an interest in LP as a non-food lignocellulosic biofuel feedstock and a carbon sequestration tool (Daystar et al., 2014; Stainback and Alavalapati, 2002; Galbe and Zacchi, 2002; Frederick et al., 2008).

Improvement of LP through classical breeding approaches is inefficient due to its relatively long generation time. However, improved understanding of the LP genome may facilitate tree improvement through marker-aided breeding and genetic engineering. The extremely large size (1C = 21.6 Gb; O'Brien et al., 1996) and complexity of LP's genome has presented challenges for its characterization (Plomion et al., 2007; Morse et al., 2009). The large genome size has primarily been attributed to an extensive accumulation of interspersed repeats, with the most significant contribution from long-terminal repeat retrotransposons (Morse et al., 2009; Ahuja and Neale, 2005; Ritland, 2012; Wegrzyn et al., 2013). In spite of the challenges, conifer genomic resources such as ESTs, full length cDNAs, SNPs, fosmid sequences, and transcriptome and protein profiling data are becoming increasingly available (Ritland, 2012; Wegrzyn et al., 2013; Kirst et al., 2003; Lorenz et al., 2006; Cairney et al., 2006; Bérubé et al., 2007; Lorenz et al., 2012; Dauwe et al., 2011; Pavy et al., 2008). To accelerate LP genomic research, we constructed a large bacterial artificial chromosome (BAC) library (> 1.8 million individual clones) for the LP genotype 7-56 (Magbanua et al., 2011). More recently, second-generation DNA sequencing and novel assembly strategies were used to prepare a high-density gene map and a draft genome assembly of LP (Wegrzyn et al., 2014; Neale et al., 2014; Neves et al., 2014; Zimin et al., 2014, 2017). As of March 15, 2018, the draft genome for LP is still highly fragmented (1.76 million scaffolds; contig N50 = 28,106 nt; https://www.ncbi.nlm.nih.gov/assembly/GCA_000404065.3/) and gene information remains largely incomplete.

The objectives of the present study were to characterize the LP genome as a whole using Cot analysis and to structurally annotate 100 LP BAC clones that were selected from our 1.8 million clone BAC library in order to explore pine gene distribution and identify genes responsible for LP's adaptive and economic traits. The Cot analysis indicates that the pine genome is 80% repetitive DNA and is characterized by a large proportion of highly diverged repeats, an observation that concurs with results from the LP genome project (Neale et al., 2014). By screening macroarrays representing a portion of the LP BAC library with overgo probes designed from cDNA sequences of wood development genes, we selected 50 "targeted" clones, each identified using a different cDNA-based probe. Additionally, we selected 25 clones based on their low-levels of repetitive DNA (as determined by macroarray screening) and 25 clones at random. The 100 BACs were Sanger sequenced, assembled, and annotated. BLASTn comparison of the cDNAs from which overgo probes were designed and BAC sequences indicated that 80% of the targeted BACs contained either all or some (22–100% query coverage) of the cDNA sequence used to target them. Using a MAKER pipeline including several gene prediction programs and cDNA support led to discovery and characterization of ten known wood formation genes and identification of an additional 144 gene models. Of note, another research group independently analyzed the same BAC sequences (Wegrzyn et al., 2013), but reported only one gene in the 100 BACs. Thus, our analysis represents an improvement on a previous effort, and provides insight into gene abundance and distribution in the LP genome. The results of this study provide insight into the overall structure of the pine genome while generating gene data that will facilitate molecular investigations of mechanisms governing wood formation.

2. Methods

2.1. Cot analysis

Young needles of LP were obtained from grafted ramets of the clone "7-56" (Ralph et al., 1997). The needles, kindly provided by International Paper, were sent to the Peterson lab wrapped in wet paper towels and sealed in plastic bags on wet ice (4 °C). Nuclear DNA isolation was performed as described previously (Peterson et al., 1997) with minor modifications as detailed in Additional File 1-Sections A–B. The DNA was evaluated using spectrophotometry where A_{260} was used to estimate DNA quantity and A_{260}/A_{280} was used to estimate purity with regard to protein (all values were adjusted for light scatter at A_{320}). The A_{260}/A_{280} ratio was always > 1.8. The purity of isolated DNA was also examined by digestion with *Hind*III (New England Biolabs) at 37 °C for 1 h where quality DNA exhibited complete digestion as determined using agarose (1% w/v) gel electrophoresis. Of note, we found that the needles could be stored at 4 °C for up to 3 months with no change in the nuclear DNA quality or yield.

DNA shearing, removal of metal ions from sheared DNA, and preparation of 0.5 M sodium phosphate buffer (SPB) are described in detail in Additional File 1-Sections C–D. In summary, nuclear DNA was sheared into fragments with a mean length of 450 bp using a Misonix Sonicator 3000. The sheared DNA was passed through a Chelex column to remove metal ions. Millipore Centriplus YM-30 columns were used to concentrate the sheared DNA and exchange the TE buffer with 0.5 M SPB. Some of the DNA in 0.5 M SPB was diluted with water to produce solutions with SPB concentrations of 0.12 and 0.03 M.

Aliquots of sheared DNA (25 μg/ml) in 0.03 M SPB and 0.12 M SPB were degassed and used to produce melting curves as previously described (Liu et al., 2011; Peterson et al., 1998, 2002). Cot analysis was performed as previously described (Liu et al., 2011). With the exception of a few of the highest Cot points (see below), each Cot point was prepared using the following steps:

(a) A 100 μg aliquot of sheared pine DNA (450 bp fragments) dissolved in 0.5, 0.12, or 0.03 M SPB was placed in a PCR tube. The DNA concentration of the sample, as determined using an Agilent 8453 spectrophotometer, was converted into moles of nucleotides per liter.
(b) A Cot value was chosen for the sample based upon its relative DNA concentration and the buffer in which it was dissolved. The reassociation time in seconds (t) for the Cot value was determined using the formula $t = \mathrm{Cot}/C{\cdot}B$ where C is the nucleotide concentration in moles per liter and B is a buffer factor that accounts for the effect of buffer cation concentration on reassociation. The values of B for 0.03, 0.12, and 0.5 M SPB are 0.0133, 1, and 5.8157, respectively (Britten et al., 1974). Reassociation times were selected so that no sample was incubated for < 1 min or > 72 h to reach its designated Cot. Cot values of DNA samples in SPB ranged from 10^{-5} to 200,000 M·s.
(c) A tube was placed into an MJ PTC-100 thermocycler with a "hot bonnet" lid. The tube was heated to 95 °C for 10 min and then immediately cooled to a temperature 25 °C below the melting

temperature for DNA in that SPB (see above) as is standard in Cot analyses (Peterson et al., 1998, 2002).

(d) The sample was allowed to reassociate until the target Cot was reached. The sample was then immediately diluted in 10–100 volumes of 0.03 M SPB to effectively stop reassociation.

(e) The diluted DNA was loaded onto a 1.2 ml HAP column prepared in a 3 ml syringe barrel. Single-stranded DNA was eluted by adding 0.12 M SPB to the column. Double-stranded DNA was eluted by adding 0.5 M SPB. Elution of single- and double-stranded fractions was monitored by spectrophotometry (A_{260}) as described previously (Liu et al., 2011).

(f) For a particular Cot value, the fraction of DNA that had reassociated was determined as detailed previously (Peterson et al., 1998).

The genome of LP is extremely large making it is nearly impossible to reach some of the highest Cot points (> 20,000 M·s) using standard techniques. Thus, we employed the ASE buffer technique to perform reassociation for Cot values > 20,000 M·s (see Additional File 1-Section F). The seven Cot points prepared using the ASE technique were 20,000, 50,000, 100,000, 250,000 (two replicates), 500,000, and 1,000,000 M·s. Reassociation was halted using 100-fold dilution in 0.03 M SPB. All other steps of fractionation were performed as described for SPB.

A best-fit Cot curve was generated from Cot data using CotQuest (Bunge et al., 2009). Highly-repetitive and moderately-repetitive Cot components were isolated as described previously (Peterson et al., 2002).

2.2. BAC library screening

The LP BAC library we constructed was utilized for this study. The preparation of the 4 × 4 macroarrays (each containing 18,432 double-spotted clones) and the screening/hybridization procedure were performed as described elsewhere (Magbanua et al., 2011). One hundred BAC clones (Table 1) were selected according to the following scheme:

(1) Targeted BAC clones – BAC macroarrays representing approximately 2.3 genome equivalents (2.3× = thirty 4 × 4 macroarrays) of LP DNA were screened with 73 overgo probes designed from cDNAs believed to represent genes associated with carbon metabolism, wood development, gene regulation (transcription factors), intracellular signaling, and/or disease resistance. Screening a 2.3× library affords roughly 90% probability of obtaining any sequence of interest (Plomion et al., 2007). Each overgo probe (consisting of a 22 bp forward and a 22 bp reverse sequence sharing an 8 bp overlap) was designed from a 36 bp cDNA region with a GC content of approximately 50%. Overgos were labeled with ^{32}P-dATP and ^{32}P-dCTP. The cDNA molecules, their GenBank accession numbers, and the overgo sequences designed from them are presented in Additional File 2-Worksheet A.
   a. Primary screen – Macroarrays were subjected to an initial screen which was split into two batches; batch 1 was performed with 25 overgos while batch 2 was performed using the remaining 48 overgos. The screening strategy employed in screening batch 1 is described in Additional File 1-Section G. A similar strategy was used for batch 2 except that the pooling pattern was in a 6 × 8 array with fourteen probe pools. Plate and well coordinates of positive BAC clones were determined from BAC macroarray hybridization patterns using manual deconvolution (Additional File 1-Section H).
   b. Secondary screen – Positive BAC clones (from the primary screen) were used to inoculate wells in microtiter plates thus creating a BAC sublibrary. The sublibrary clones were used to create macroarrays, and the macroarrays were screened with overgo row and column pools; this allowed each overgo (and ostensibly the cDNA from which it was designed) to be assigned to one or more BACs (see Additional File 1-Section H).
   c. Selection of targeted clones for sequencing – All positive clones identified in the secondary screen were analyzed using *Not*I digestion (to release inserts from the BAC vector) and pulsed-field electrophoresis to verify that the clones contained insert DNA (see Additional File 1-Section H). Of the overgo probes, 42 of 73 (57.5%) hybridized to between one and seven clones, 28.8% (21/73) hybridized to more than seven clones, and ten did not produce a hybridization signal. As only a part of the LP BAC library was screened, we anticipated that some overgos would not find corresponding BAC clones. In cases where more than one clone was identified for a probe, we selected the clone with the highest signal intensity. If signal intensity was roughly equivalent, the clone with the largest insert was selected. Fifty of the BAC clones, each identified by a different overgo probe, were then selected for sequencing (Additional File 2-Worksheet B).

(2) Low-copy sequence clones - The first BAC macroarray prepared from the library (which contains 18,432 BAC clones) was screened using Cot-fractionated highly-repetitive and moderately-repetitive DNA components (see above). Briefly, 100 ng from the highly- and moderately-repetitive components were separately labeled with ^{32}P-dCTP, pooled, and used as to probe the macroarray. Twenty-five clones with inserts > 100 kb that showed little or no obvious hybridization to the repetitive DNA were chosen for sequencing. Additionally, the chosen “low-repeat” BAC clones were not recognized by the cDNA-based overgo probes (see Additional File 1-Section H).

(3) Random clones - Clones selected at random from the BAC library were examined by pulsed-field gel electrophoresis, and the 25 with the largest insert sizes (that had not been selected as targeted or low-copy clones) were chosen for sequencing.

2.3. BAC sequencing

BAC DNA was isolated from a single bacterial colony and purified using a Qiagen MaxiPrep column (Qiagen, Valencia, CA). DNA was sheared to yield 3–4 kb fragments using Adaptive Focused Acoustics technology (Covaris, Woburn, MA) that were cloned into the plasmid vector pIK96 as previously described (Ferris et al., 2010). Universal primers and BigDye Terminator Chemistry (Applied Biosystems, Grand Island, NY) were used for Sanger sequencing of randomly selected plasmid subclones to a depth of 10×.

The Phred/Phrap/Consed suite of programs was used to assemble and edit the sequence with following assembly parameters: -vector_bound 20 -minmatch 30 -maxmatch 55 -minscore 55 (Ewing et al., 1998; Ewing and Green, 1998; Gordon et al., 1998). After manual inspection of the assembled sequences, finishing was performed by re-sequencing plasmid subclones and by gene-walking on plasmid subclones or the BAC clone using custom primers. All finishing reactions were performed using dGTP BigDye Terminator Chemistry (Applied Biosystems). Finished clones contained no gaps and were estimated to contain less than one error per 10,000 bp.

Table 1
Summary of pine BAC clones sequenced.

Number of targeted BAC sequences	49[a]
Number of low-repeat BAC sequences	25
Number of random BAC sequences	25
Total base count (bp)	11,597,749
Minimum BAC length (bp)	8288
Maximum BAC length (bp)	172,161
Mean BAC length (bp)	115,977
GC content	38%

[a] While 50 targeted BACs were sequenced and assembled, one of the targeted BACs (AC241291.1) was found to contain fungal DNA and was eliminated from further study.

A list of the BAC clones that were selected for sequencing and their corresponding GenBank accession numbers are presented in Additional File 2-Worksheet B.

2.4. Success of targeting

As described above, overgo probes were designed based upon cDNA sequences representing genes ostensibly involved in carbon metabolism, wood formation, disease resistance, etc. To determine the relative success of BAC targeting, blastn was used to align each overgo-targeted BAC with its corresponding cDNA. The query coverage, E value, max bit score, total bit score, and identity values were recorded for each successful cDNA-BAC alignment.

2.5. Repeat identification

The homology-based repeat identification tools RepeatMasker (Smit et al., 2013) (RMLib: 20140131 & Dfam: 1.2) and RepeatRunner (Smith et al., 2007), powered by the sequence search engine AB-*BLAST* (an updated form of WU-BLAST; Tarailo-Graovac and Chen, 2009), were used to identify repeats in the LP BACs. RepeatMasker was used to compare BAC sequences with the Viridiplantae section of the Repbase repeat database (release 05-10-2010) while RepeatRunner used its default te_protein.fasta file as a blastx database. Default parameters were used with all programs. RepeatRunner was used to compare a library of known mobile element proteins to those regions of LP BACs not recognized as repeats by RepeatMasker.

2.6. Gene identification

BAC sequences were annotated using the genome annotation pipeline MAKER (version 2.2 installed in high performance computing clusters at Mississippi State University), which identifies and masks out repeats using RepeatMasker and RepeatRunner (see above), aligns EST and protein homology evidence to a target genome using BLAST and Exonerate (Slater and Birney, 2005), produces ab initio gene predictions using appropriate scripts, and automatically integrates all these data to produce final gene models with evidence-based quality statistics (Yandell and Ence, 2012; Holt and Yandell, 2011; Campbell et al., 2014). LP expressed sequences (EST, cDNA, and mRNA) from GenBank were used as direct evidence for exon prediction. Annotated coding and protein sequences available for LP (http://pinegenome.org/pinerefseq/) were also utilized as further evidence. Ab initio gene predictions were made inside of MAKER by Augustus (Stanke et al., 2004). Furthermore, FGENESH (Solovyev et al., 2006) and Genemark (Lomsadze et al., 2005) ab intio prediction programs were incorporated exogenously into MAKER using the pred_gff parameter. These programs have not been optimized to search for conifer genes; indeed, at the time of the analysis the choices of plants were limited to *Arabidopsis* and wheat, both of which are separated from the gymnosperms (including LP) by 310 million years (Li et al., 2015). For better or worse, the gene finding parameters for wheat (*Triticum aestivum*) were used in all three ab intio predictors. Final annotations generated using MAKER were visualized using the Artemis genome viewer (Rutherford et al., 2000) for manual curation. To identify homologies, we conducted blastn alignment to the NCBI GenBank non-redundant nucleotide database, using an e-value threshold of 1e-10. Ten complete genes were identified and characterized; the genes are annotated in their corresponding BAC submissions, and they have their own unique GenBank protein accession numbers (AHX74218.2-AHX74225.2, AHX59161.2, and ANA07245.1).

2.7. Comparison of BAC sequences to draft LP genome

Our 100 LP BAC sequences were compared to the draft LP genome using blastn (v2.2.30) with default filtering and alignment parameters. Additional File 2-Worksheet B shows each BAC accession (query), the LP draft genome scaffold id (subject) with the highest query coverage per subject, and the query coverage per subject for each BAC sequence.

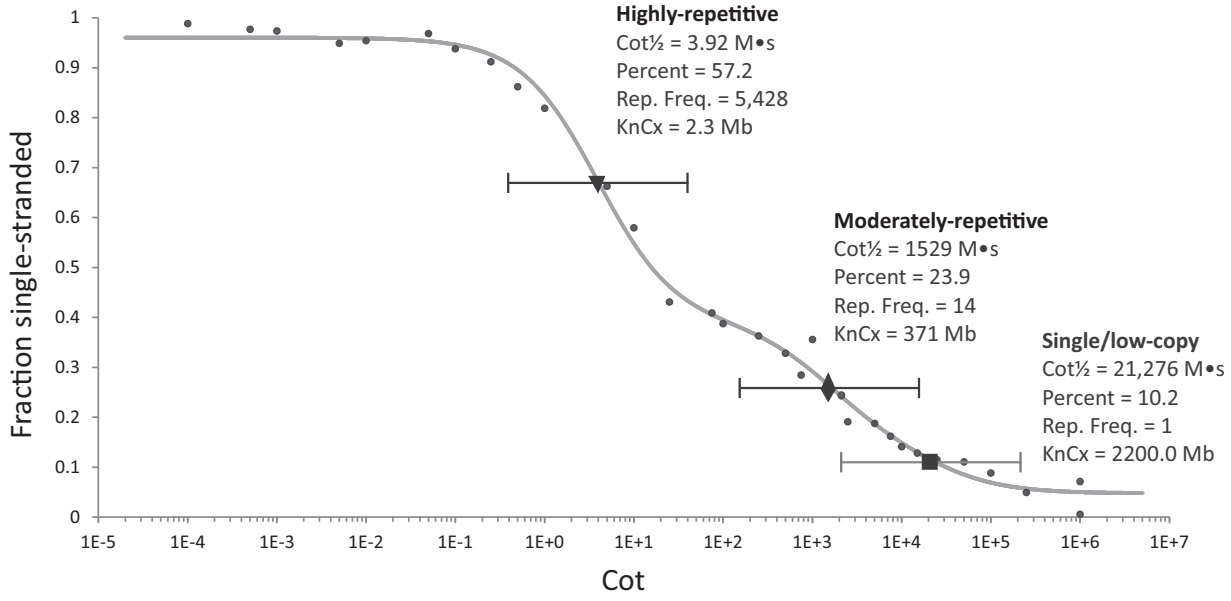

Fig. 1. Cot curve for loblolly pine. All Cot analysis results are presented in the figure. A least-squares curve was fit through the data points (circles) using the CotQuest program (Bunge et al., 2009). The curve consists of highly repetitive (HR), moderately repetitive (MR) and single/low copy (SL) components, and a triangle, a diamond, and a square mark the Cot½ values of the HR, MR, and SL components, respectively. The brackets centered at a particular Cot½ marker show the "two Cot decade region" in which 80% of the sequences in that component will renature (Peterson et al., 2002). Information for each Cot component is shown to the right of its position on the Cot curve. Rep. Freq. = repetition frequency; KnCx = kinetic complexity, the amount of novel DNA sequence in a particular component. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Results and discussion

3.1. Cot analysis

Melting curves were generated for sheared LP DNA in 0.03, 0.12, and 0.5 M SPB, and melting temperatures (Tm) for DNA in each buffer were determined using first-derivative analysis. The Tm for LP DNA in 0.03, 0.12, and 0.5 M SPB are 75.7, 83.8, and 91.8 °C, respectively. For DNA dissolved in buffers with a monovalent cation concentration (M_{mvc}) between 0.01 and 0.2 M, the GC content of the DNA can be calculated using the formula %GC = 2.44 (Tm-81.5-16.6 logM_{mvc}) (Mandel and Marmur, 1968). The LP DNA samples in 0.03 M SPB (Na^+ = 0.045 M) and 0.12 M SPB (Na^+ = 0.18 M) result in %GC estimates of 40.4% and 42.9%, respectively (average = 41.7%). The LP GC content of 41.7% as determined by DNA melting is similar to the GC percentage for LP determined by draft genome sequencing (38.2%; Neale et al., 2014) and the GC contents determined for five other pines by flow cytometry (39–40%; Bogunic et al., 2003).

The CotQuest nonlinear regression model that provided the best fit of the renaturation kinetics data was a three-component fit where the reassociation rate (*k*) of the slowest reassociating component was fixed, based on the genome size of LP (see Peterson, 2005 for details). The CotQuest program detected one Cot point that was a statistically valid outlier. However, removal of the outlier from the data had no impact on the best-curve model chosen by the program or on the biological values derived from the curve. As per CotQuest user instructions (Bunge et al., 2009), the complete data set should be used if outlier removal has no effect on model selection. The CotQuest program also has the ability to generate Cot curves using both "Gauss" and "Marquardt" numerical search algorithms. Both algorithms produced essentially identical results for LP. The Marquardt fit of the complete data set was chosen to characterize the LP genome. The LP Cot curve and the major biological findings obtained from curve analysis are shown in Fig. 1. Of note, the curve is composed of highly repetitive (HR), moderately repetitive (MR), and single/low-copy (SL) components accounting for 57.2, 23.9, and 10.2% of the genome, respectively. Because the reassociation rate of the SL component was fixed based upon the genome size, the resulting curve cannot be used to estimate the LP genome size. The SL component supposedly has a repetition frequency of 1. To determine the repetition frequency of the MR and HR components, the Cot½ value of the SL component was divided by the Cot½ values obtained for the MR and HR components, respectively (Liu et al., 2011). The mean predicted repetition frequencies of sequences in the MR and HR components are 14 and 5428, respectively. Fold back DNA and unreassociated (ostensibly damaged) DNA, features of all Cot curves (see Peterson, 2005 for review), accounted for 1.2% and 4.74% of the genomic DNA, respectively.

As with other conifers (see MacKay et al., 2012), the LP genome is extremely large. Collectively, HR and MR components account for 81.1% of the genome. Our estimate of repeat content agrees with previous reassociation kinetics studies of four other closely related conifers – specifically, *Pinus lambertiana*, *P. resinosa*, *P. banksiana*, and *Picea glauca* – where the results indicated that about 70–80% of each of these genomes is repetitive DNA (Rake et al., 1980). Likewise, analysis of the draft genome sequence of LP led to the conclusion that the genome of LP was 82% repetitive (Neale et al., 2014). Of note, our previous Cot analysis of the more distantly-related conifer, *Taxodium distichum* (bald cypress), which has the smallest conifer genome (9.7 Gb; Hizume et al., 2001), was found to possess a higher proportion of repetitive DNA (90%; Liu et al., 2011).

Our Cot analysis indicates that 10.2% of the LP genome is single/low-copy DNA. Cot curve-based SL estimates for *Pinus lambertiana*, *P. resinosa*, *P. banksiana*, and *Picea glauca* are 2–3 times higher (20–30%; Rake et al., 1980) than the percentage we obtained for *P. taeda*. However, there are several reasons to believe that the previously published SL estimates (Rake et al., 1980) may be flawed. First of all, it appears that the authors "normalized" their Cot curves by distributing the fold back and unreassociated DNA fractions, common to all Cot curves prepared using hydroxyapatite chromatography, between the kinetic components. In other words, the sum percentage of DNA in the kinetic components for each curve (in Rake et al., 1980) add up to 100% reassociation, although in reality reassociation never actually starts at 0% or ends at 100% (Britten et al., 1974; Peterson et al., 2002). Such "normalization" of Cot curves ostensibly inflates the contribution of some or all components and is not recommended as the natures of fold back and unreassociated fractions have not been well studied (Peterson et al., 2002). Additionally, the Cot components elucidated by Rake et al. (1980) were determined by eyeball identification of linear regions within the Cot curve, a practice that is more subjective and likely to lead to error than utilization of a curve-fitting program designed for analyzing Cot data (e.g., NNNBAT, Pearson et al., 1977; and CotQuest, Bunge et al., 2009). Assuming that our 10.2% estimate of SL DNA is correct, the SL component of pine (21.6 Gb × 0.102 = 2.2 Gb) is still 14 times larger than the *Arabidopsis thaliana* genome (156 Mb; Bennett and Leitch, 2005). As LP has no more than two times as many genes as *Arabidopsis* (Neale et al., 2014; Swarbreck et al., 2008), it is clear that much of the SL DNA in LP is non-coding DNA (see Section 3.7 *Gene/repeat-poor DNA* below).

3.2. Contamination and potential chimeric inserts

BLASTn analysis was conducted on all the BACs as part of the characterization process. Of note, one randomly selected BAC (AC241292.1) was found to contain at least 94,714 bp of DNA that showed significant homology to the genome of the conifer root-rot fungus *Heterobasidion irregulare*, and hence this BAC was eliminated from additional analyses. A second targeted BAC (AC241334.1) was found to contain about 1000 bp of conifer chloroplast DNA near its center. However, the remainder of the BAC appeared to contain LP nuclear DNA as it showed its strongest BLAST hits to 20 of our other LP BACs and a *Picea glauca* mRNA sequence (total bit scores between 2171 and 206,100) and appeared to have 65% of its length in an LP draft genome scaffold. The clone may be chimeric, although it is possible (though less likely) that a piece of chloroplast DNA was integrated into the LP 7-56 genome and detected by BAC sequencing. This BAC was included in further analyses.

3.3. Targeted BACs

BLASTn analysis indicates that of the 50 targeted BAC clones, 40 possess a region(s) that shows significant alignment to the cDNA from which the overgo was designed (E values range from 0 to 6.00E-42, and total bit scores range from 387 to 6366). Thus our success rate at obtaining BACs with high level identity to the cDNAs used to target them was 80%. Twenty-nine of the 50 targeted BACs (58%) contained what appeared to be an intact or nearly intact copy of the target cDNA while 11 (22%) contained portions of coding sequence, but appeared to lack a full-length target gene. Ten of the 50 clones were false positives (i.e., they did not contain the target cDNA or the overgo probe). As discussed above, one of these false positive BACs contained fungal DNA and was eliminated from further evaluation. Of the remaining nine, we discovered in post-sequencing analysis that one clone (GenBank Acc. AC241313.1) was targeted using an overgo of unknown origin with no sequence similarity to any of the cDNAs, LP BACs, or any sequence in GenBank. We are certain human error was responsible for design and inclusion of the mystery overgo in the experiments. While we are not certain why the false positive percentage (20%) was relatively high, we do note that deconvoluting macroarray hybridization images can be difficult considering the high density of clones (18,432 double-spotted clones) per macroarray, i.e., human error is very possible.

The BLASTn results are summarized in Additional File 2-Worksheet C.

3.4. Repeat analysis

The pine genome is rich in repeats, and there are a number of excellent papers discussing the repeat sequence content of LP (Morse et al., 2009; Kovach et al., 2010; Neale et al., 2014), including an independent repeat analysis of the BAC clones we sequenced (see Wegrzyn et al., 2013). We did, however, use RepeatMasker and RepeatRunner to mask out known repeats prior to gene model identification. A glance at the masking results suggest that 16.49% of the nucleotides in the complete 99 BAC set belong to known plant repeat groups. This low percentage of repeats is not particularly surprising as Wegrzyn et al. (2013) noted that in their attempts to characterize repeats in pine DNA (including the BACs in this study), homology-based methods yielded an estimated repeat content of only 27% for LP; it was not until ab initio repeat identification approaches were utilized that the repeat percentage rose above 80%. The focus of Wegrzyn et al. (2013) was on repeats, and their use of multiple ab initio repeat identification programs allowed them to show that LP contains many sequences that contain repeat structural features – suggesting that they are derived from repeat sequences – that are not repetitive (i.e., are found in a single-copy per 1C genome) or only found in two or three copies. Of the 6270 full-length transposable element families they discovered, 82% were annotated as single-copy (Wegrzyn et al., 2013; Neale et al., 2014; Wegrzyn et al., 2014). We did not screen the BACs with the 6270 transposable element family representatives, in part because DNA sequences with mobile element-like features that are found only once in a genome are arguably not repeats. More importantly, the analysis of Wegrzyn et al. was extremely thorough, and we were not looking to replicate their analyses. Instead our focus was on genes which were not the primary focus of their paper (see below).

Of interest, the repeat content of the different pine BAC types (low-copy, random, and targeted) have statistically indistinguishable mean values (Table 2; see Additional File 2-Worksheet F for statistical tests). This finding could mean that low-repeat BACs (chosen based on their relative lack of hybridization with Cot-isolated repeats) do not actually have fewer repeats than random and targeted clones. However, because so many pine repeats are found in low-copy numbers and/or have not been described, it would require considerably more analysis of the BACs to determine if the different BAC types truly are similar with regard to repeat content. While Wegryzn did a detailed evaluation of the repeat content of the BACs, they considered the BACs as a single group, so their results cannot be used to draw conclusions about the repeat distributions in the different BAC types.

Our admittedly limited homology-based results do agree with previous studies showing that LTR retroelements are the dominant repeat type in LP with members of the *Gypsy* subclass being more common than those from the *Copia* subclass (Table 3). LTRs retrotransposons are one of the primary agents underlying genome expansions in both gymnosperms and angiosperms (Morse et al., 2009; Ahuja and Neale, 2005; Kovach et al., 2010).

3.5. Comparison of BAC sequences with draft LP genome sequence

The 99 BAC clones were compared to the draft LP genome sequence using BLASTn. Parameters were set so that we could get a rough estimate of how much of each BAC clone was found within the LP draft genome (Additional File 2-Worksheet B). Nine of the clones showed > 90% sequence identity to scaffolds/contigs in the draft genome. Five showed 80–89% sequence identity, sixteen exhibited 70–79%, twenty-five had 60–69%, twenty-two showed 50–59%, twelve exhibited 40–49%, and ten had 30–39%. The minimum query coverage was 33% while the maximum value was 99%.

3.6. Genes and gene models

The 99 BAC clones containing LP DNA were analyzed using combinations of similarity and ab initio gene prediction approaches. These analyses resulted in discovery of 154 gene models (GMs) – see Additional File 2-Worksheet E. Of these, ten were full-length genes known to be involved in wood formation and carbon metabolism. These "verified" genes, which were all found within the targeted BAC set, include two GDP-mannose pyrophosphorylases (i.e. *GMP2* and *GMP1*), caffeoyl-CoA *O*-methyltransferase (*CCoAOMT*), laccase (*LAC8*), Korrigan endoglucanase (*KOR1*), two cellulose synthases (*CesA1* and *CesA2*), phenylalanine ammonia lyase (*PAL*), sucrose synthase (*SuSy1*), and a MYB transcription factor (*MYB8*). The predicted structures of the verified genes are shown in Fig. 2, and each is discussed in more detail below.

The remaining 144 GMs represent a combination of possible coding genes (one or more open reading frames, ORFs, flanked by a transcriptional stop and start codon) and likely pseudogenes (sequences with homology to known genes but lacking stop and/or start codons) (Table 3 and Additional File 2-Worksheet D). Despite the temptation to do so, we refrained from defining each GM as a possible coding gene or pseudogene as it is well known that even a GM that has all the expected parts of a protein coding gene may never be translated, while a sequence without all the features of a protein-coding gene may be transcribed and/or translated.

The 154 GMs (including the 10 characterized genes from targeted BACs) had roughly equal distributions in the targeted and random BAC clones (15.85 GM/Mb and 13.18 GM/Mb, respectively). However, the low-repeat BACs contained 2–2.5-fold fewer GMs (6.44 GM/Mb). Interestingly, if the total length of GMs is presented as a fraction of BAC length, the differences between the BAC types are more pronounced. As shown in Table 3, the targeted and random BACs contain 5.3 and 3.0 times more of their length in GMs than do low-copy BACs. The mean GM length in targeted BACs (2063 bp) is significantly different from the mean GM lengths for random (1387 bp) and low-copy (952 bp) BACs while the difference between the low-copy and random BAC GM length means is not statistically significant ($\alpha = 0.05$) (see Additional File 2-Worksheet F for statistical test results).

In an initial, independent analysis of the 100 pine BAC clones produced by our research team, another research group (Wegrzyn et al., 2013) found only one gene (which they did not identify by name). Our analysis indicates that the BAC clones are significantly richer in genes and GMs than previously suggested. In our experience, even minor changes in script parameters can make enormous differences in the number of gene models identified by bioinformatics programs. Such minor differences could account for the discrepancy between our results and those of Wegrzyn and her colleagues.

Table 2
Repeat distributions (% total length) in the pine BAC sequences based on RepeatMasker & RepeatRunner analyses.

		All BACs (n = 99)	Low-repeat (n = 25)	Random (n = 25)	Targeted (n = 49)
Class I - retroelements	LINEs	0.26	0.64	0.17	0.18
	LTR-Copia	4.75	3.09	4.82	5.31
	LTR-Gypsy	7.24	7.31	7.60	6.99
	LTR-uncertain	0.50	0.40	0.75	0.40
	Other retroelements	0.34	0.17	0.51	0.30
Class II - DNA transposons	TIR	0.15	0.37	0.10	0.10
	Helitron	0.00	0.01	0.00	0.00
	Maverick	0.00	0.01	0.00	0.00
	Simple repeat[a]	2.21	3.57	2.01	1.83
	Unspecified	1.03	0.94	1.07	1.04
	Totals	16.49	16.51	17.03	16.16

[a] Simple repeat includes low complexity sequences and satellite DNAs.

Table 3
Gene model (GM) information by BAC type.

BAC type	Number of GMs identified	Cumulative BAC assembly length (bp)	Total bp in GMs	Mean GMs/Mb	Mean percent length in GMs	Mean GM length (bp)	Mean exon number per GM
Low-repeat	14	2,175,369	13,348	7.08	0.62	953.43	2.14
Random	46	3,490,770	63,835	12.61	1.63	1387.72	2.24
Targeted	94	5,931,610	194,026	15.19	3.26	2064.11	3.15

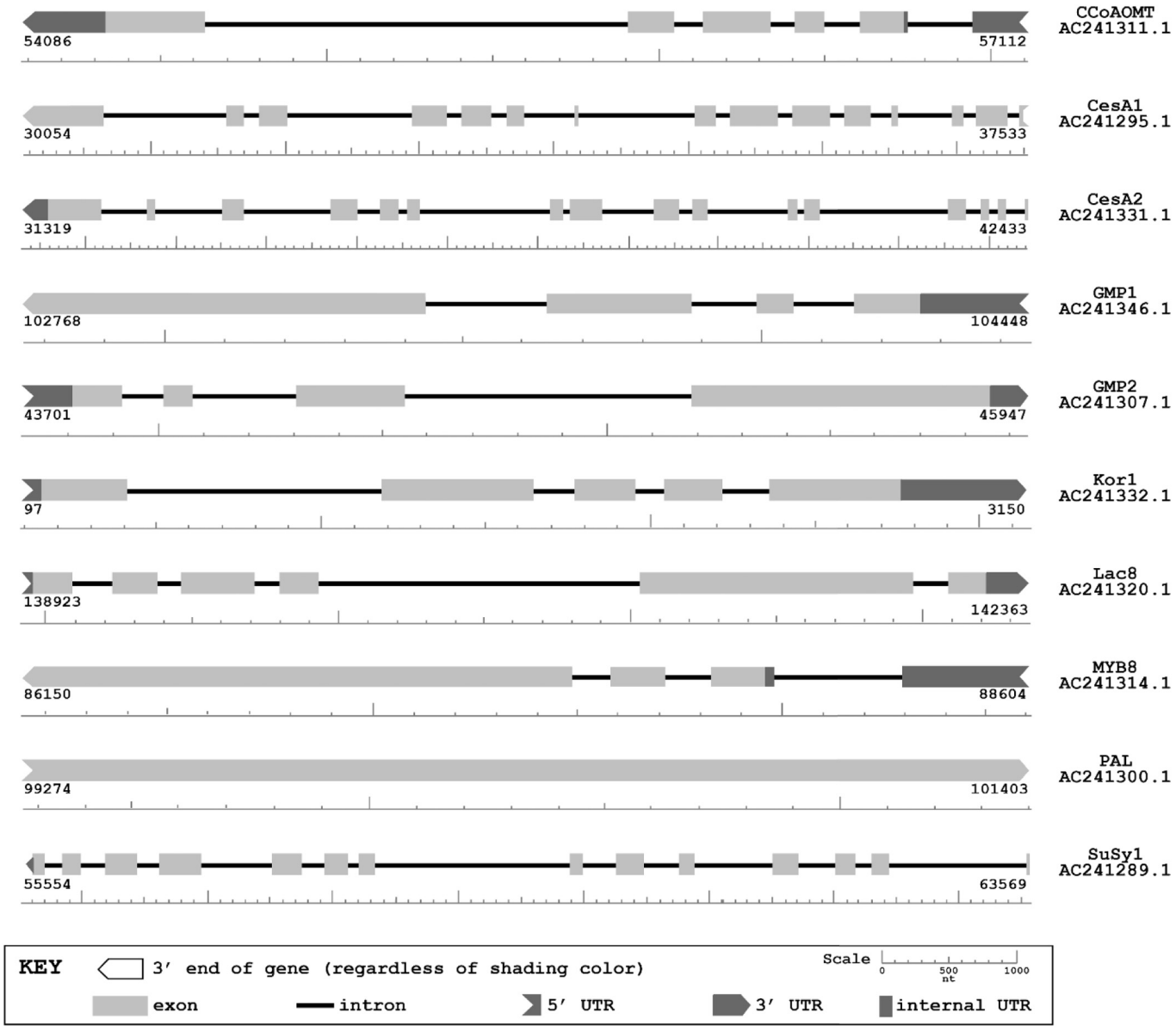

Fig. 2. Intron-exon structures of ten characterized genes related to loblolly pine wood formation. The acronym for a gene and the accession number for the BAC in which the gene is found is to the right of its model. The numbers at each end of a gene indicate its position within its BAC clone.

3.7. Gene/repeat-poor DNA

The Cot analysis indicates that the single/low copy component of the LP genome is 14 times the size of the *Arabidopsis* genome. However, the number and size of genes in pines is not 14 times greater than that of *Arabidopsis*; LP has an estimated gene model number of 50,172 while *Arabidopsis* has 37,898 genes and 4843 pseudogenes (https://www.ncbi.nlm.nih.gov/genome/4). While LP has some genes with exceedingly long introns (e.g., the longest known LP intron is 319 kb) and high average intron lengths (2.7 kb), intron size is insufficient to explain the large amount of single-copy DNA in pine (especially when one considers that LP introns are enriched for transposons) (Wegrzyn et al., 2014). Consequently, it appears that much of the LP single/low-copy DNA is both gene- and repeat-poor.

What is the origin of gene/repeat-poor DNA in LP? Because polyploidy events in conifers are extremely limited and exceedingly ancient (Li et al., 2015), it is unlikely that large-scale deterioration of duplicated genes is responsible for gene/repeat-poor DNA. A much more likely explanation is that LP experienced a number of mobile element amplification events followed by extended periods of repeat sequence

divergence, a hypothesis that has been propounded on numerous occasions (Morse et al., 2009; Wegrzyn et al., 2013; Neale et al., 2014). In an analysis of completed angiosperm genomes, El Baidouri and Panaud (2013) showed that mobile element amplification typically occurs in short duration evolutionary *bursts* usually involving one mobile element family. Each burst is typically followed by a series of events that result in the elimination of many/most repeat copies; such events include deletions, unequal recombination, and deletion-biased double strand break repair (El Baidouri and Panaud, 2013; Schubert and Vu, 2016). However, draft genome analyses of conifers (LP, *Picea glauca*, and *Picea abies* – Neale et al., 2014; Birol et al., 2013; Nystedt et al., 2013) and other gymnosperms (*Gnetum montanum* and *Ginkgo biloba* – Wan et al., 2018; Guan et al., 2016) indicate that this clade does not eliminate repeats in an efficient manner (see Pellicer et al., 2018 for review) leading to mobile element accumulation. In support of this notion, Cossu et al. (2017) found that unequal recombination between LTR retroelements, which tends to result in repeat elimination, is less frequent in conifers than in small genome angiosperms. With this said, gymnosperms are not the only plant genomes that carry significant amounts of repetitive DNA that has diverged to the point of being no longer repetitive. Members of the lily genus *Fritillaria* (1C = 30–100 Gb; Kelly et al., 2015), whose genomes are larger than conifer genomes, also appear to have ancient, highly-diverged repeats. While one might conclude that inefficient repeat elimination is limited to species with massive genome sizes, *Amborella trichopoda* (1C = 0.87 Gb) also has a genome characterized by ancient, highly diverged repeats (Amborella Genome Project, 2013).

All conifers have large genomes and amazingly stable chromosome numbers (e.g., all but three of the 156 Pinaceae species for which chromosomes have been counted have a chromosome number of $n = 12$) (Murray et al., 2012; Pellicer et al., 2018), it is logical to think that many of the large retroelement bursts that contributed to gene/repeat-poor DNA occurred early in conifer evolutionary history. An ancient origin of the sequences in gene/repeat-poor DNA would be consistent with tremendous sequence heterogeneity.

3.8. Carbon metabolism and wood-formation genes

We identified ten full-length genes that code for proteins involved in carbon metabolism and wood formation. Below is a summary of these genes and their roles in LP:

(A) *Cellulose synthase* (*CesA*) - Cellulose, the most abundant biopolymer on earth, is a major component of wood (Nairn et al., 2008). Cellulose is synthesized at the plasma membrane by large membrane-localized protein complexes with each complex containing several structurally similar cellulose synthase (CesA) subunits encoded by different members of the *CesA* gene superfamily and the cellulose synthase-like (*Csl*) gene family (Harris and DeBolt, 2010; Brown, 1996; Saxena and Brown, 2005). Since the first *CesA* gene was characterized in cotton, more CesA and Csl gene sequences have been identified in various plant species including ten *CesA* sequences in *Arabidopsis*, forty-five *CesA/Csl* sequences in rice (*Oryza sativa*), and twelve *CesA* sequences in corn (*Zea mays*) (Pear et al., 1996; Wang et al., 2010; Appenzeller et al., 2004; Endler and Persson, 2011). Further, recent research progress in wood formation has led to the identification of *CesA* sequences in tree species of great economic value, including seven and eighteen *CesA* genes in *Populus tremuloides* and *Populus trichocarpa* respectively, six *CesA* genes from *Eucalyptus grandis*, and ten *CesA* genes in LP (Nairn et al., 2008; Palle et al., 2011; Joshi et al., 2004; Djerbi et al., 2005; Ranik and Myburg, 2006). In our analysis of the BAC clones, using a combination of similarity and ab initio gene prediction approaches, we identified and characterized two *CesA* genes (i.e. *CesA1* and *CesA2*) which are 7.4 kb and 11.1 kb in length respectively. Sequence alignment showed that both genes have 99% sequence identity (E-value = 0) to the *CesA1* and *CesA2* sequences previously identified in EST collection of LP (Nairn and Haselkorn, 2005). Also, *CesA1* shared 75–99% (E-value = 0) sequence identity to *CesA* homologs of *Pinus radiata*, *Pinus pinaster*, and *Cunninghamia lanceolata*. LP *CesA1* is composed of fifteen exons separated by fourteen introns and LP *CesA2* is composed of sixteen exons separated by fifteen introns (Fig. 2). Intron-exon structure of LP *CesA* genes is quite similar to the poplar (*Populus tremuloides*) *CesA* homologs; both poplar and LP *CesA* genes range from 4.8–7.2 kb in size and possess 11–14 introns which suggests that these genes might be structurally and possibly functionally conserved (Joshi et al., 2004).

(B) *KORRIGAN* (*KOR*) – Apart from cellulose synthases, a membrane-anchored cellulase (an endoglucanase) involved in primary and secondary cell wall synthesis has been identified in both gymnosperms and angiosperms, including forest tree species such as white spruce (*Picea glauca*) (Maloney et al., 2012), LP (Nairn et al., 2008), and hybrid poplar (*Populus alba* × *grandidentata*) (Maloney and Mansfield, 2010). More specifically, β-1,4-endoglucanase (EC 3.2.1.4), encoded by the *KORRIGAN* (*KOR*) gene, has been linked to cellulose biosynthesis and/or deposition in plant cell walls (Mølhøj et al., 2002). We identified a *KOR* gene which is 3 kb in length sharing the highest sequence identity with the LP *KOR* (98%) and *Picea glauca KOR* (92%) at 0 E-value. The LP *KOR* gene is organized into five exons and four introns (Fig. 2); this organization is shared with *Arabidopsis*, *Picea glauca*, and *Populus tremuloides KOR* orthologs (Maloney et al., 2012; Stival Sena et al., 2014). KOR is believed to be involved in cell plate formation during cytokinesis (Zuo et al., 2000), xylem vessel development (Szyjanowicz et al., 2004), cellulose synthesis and deposition in secondary xylem of various species including *Populus* spp., indicating the important role for KOR's in wood formation (Maloney and Mansfield, 2010; Yu et al., 2014). Moreover, the rescue of the *Arabidopsis thaliana kor1-1* irregular xylem and dwarf phenotype by the expression of endogenous *KOR* from white spruce (*Picea glauca*) revealed functional conservation of *KOR* gene between gymnosperms and angiosperms (Maloney et al., 2012). However, the precise mechanism by which KOR protein is involved in cellulose synthesis remains unknown.

(C) *Sucrose synthase* (*SuSy*) – Genes encoding sucrose synthase (SuSy, EC 2.4.1.13) have been demonstrated to be involved in cellulose biosynthesis; SuSy cleaves sucrose to produce fructose and UDP-glucose, and the latter serves as a precursor for cellulose biosynthesis (Fujii et al., 2010; Guerriero et al., 2010). Overexpression of *SuSy* genes resulted in elevated cellulose synthesis (up to 6% over control levels) in poplar (Coleman et al., 2009), increased plant height and biomass in tobacco (Coleman et al., 2006), and accelerated leaf expansion and enhanced fiber production in cotton (Xu et al., 2012). Gene expression profiling studies revealed that *SuSy* transcripts were highly abundant in LP wood forming tissues along with other transcripts for the cellulose synthase complex (Nairn et al., 2008; Yang et al., 2004; Whetten et al., 2001). With emerging genome sequence data, multiple members of *SuSy* gene family have been identified and characterized in various species, including six *SuSy* genes in *Arabidopsis* (Baud et al., 2004) and rice (*Oryza sativa*) (Hirose et al., 2008), and fifteen in *Populus* (An et al., 2014). Further, EST data revealed multiple *SuSy* genes in LP (Nairn et al., 2008). During our analysis, we identified a *SuSy* gene that is 8 kb in length and shares highest sequence identity (90–99%) with LP, *Picea sitchensis*, and *Pinus pinaster SuSy* genes. The *SuSy* gene sequence from the BAC clone was interrupted by fourteen introns, which is consistent with intron-exon structures for other *SuSy* genes from *Pinus pinaster*, *Arabidopsis*, rice (*Oryza sativa*), *Populus*, and cotton (*Gossypium arboreum*) with 12–14 exons, indicating structural conservation of *SuSy* homologs (An et al., 2014; Chen et al., 2012; Seoane-Zonjic et al., 2016). Besides its role in cellulose

biosynthesis, SuSy is reported to be involved in environmental stress tolerance (Hirose et al., 2008; Geigenberger et al., 1997), starch biosynthesis (Chourey et al., 1998; Baroja-Fernández et al., 2003), and nitrogen fixation (Gordon et al., 1999; Horst et al., 2007), suggesting functional divergence of the *SuSy* gene family during evolution (Chen et al., 2012).

(D) *GDP-D-mannose pyrophosphorylase* (*GMPase*) – Besides cellulose, hemicelluloses (i.e. xyloglucans, xylans, mannans, and glucomannans) are a major component of plant cell walls (primary and secondary), and thereby contribute to wood formation (Lerouxel et al., 2006). Complex cross-linking of cellulose by hemicelluloses during growth and morphogenesis permits the cell wall to be strong, yet flexible (Scheller and Ulvskov, 2010). Glucomannans, desirable in energy feedstocks for bioethanol production, are more abundant in the secondary cell wall of LP and other softwoods, whereas xyloglucans are more abundant in angiosperm secondary cell walls (Pauly and Keegstra, 2008; Suzuki et al., 2006). The primary substrate for assembly of the glucomannan hemicelluloses in the cell walls of wood-forming tissues is GDP-D-mannose, and the polymerization reaction is catalyzed by mannan and glucomannan synthase activities encoded by *CslA* genes, which have been identified in a number of plant species (Nairn et al., 2008; Pauly and Keegstra, 2008; Liepman et al., 2007; Dhugga et al., 2004; Hazen et al., 2002). More specifically, GDP-D-mannose pyrophosphorylase (GMPase, EC 2.7.7.22), localized to the cytosol, is the enzyme that catalyzes the synthesis of GDP-D-mannose from D-mannose-1-phosphate (Conklin et al., 1999). During our analysis, two GMPase-encoding LP genes, *GMP1* and *GMP2*, which are 1.6 kb and 2.2 kb in length, respectively, were identified. Sequence alignment revealed that *GMP1* shares 100% (E-value = 0) and 95% (E-value = 0) sequence identity to LP and *Picea sitchensis* GMPase genes, respectively, whereas *GMP2* shares 99% (E-value = 0) and 92% (E = 0) sequence identity to LP and *Picea glauca* GMPase genes, respectively. *GMP1* and *GMP2* consist of four exons separated by three introns (Fig. 2) which is in agreement with the number of exons in the GMPase homolog of acerola (*Malpighia glabra*) (Badejo et al., 2008). In a comparative genomics study of LP ESTs where putative LP protein sequences were compared with those from angiosperms and a phylogeny inferred from protein sequence alignments, it was found that the two LP GMPase genes are more closely related to each other than to orthologs from other taxa (Nairn et al., 2008).

(E) *Phenylalanine ammonia lyase* (*PAL*) – Lignin, the second most abundant plant biopolymer after cellulose, is particularly abundant in wood-forming cells, such as tracheids, wood fibers, and vessel elements that undergo secondary cell-wall thickening, and is covalently linked to non-cellulosic cell-wall polysaccharides (Wagner et al., 2012; Zhong and Ye, 2009). The polymer is vital for plant fitness, vascular integrity, and pathogen defense in conifers, although it is an undesirable barrier to pulp and paper manufacturing. Some of the key molecular players involved in the lignification process has been identified and characterized in various plant species including conifers such as LP (Whetten et al., 2001) and *Pinus radiata* (Wagner et al., 2011).

Lignin is created though the dehydrogenative polymerization of *p*-hydroxycinnamyl alcohols (monolignols): *p*-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol (Zhong and Ye, 2009). Of note, *p*-coumaryl alcohol and coniferyl alcohol are the principal building blocks of conifer (softwoods) lignin, which is primarily composed of *p*-hydroxyphenyl (H) and guaiacyl (G) units, and lacks the sinapyl alcohol-derived syringyl (S) units that are commonly found in hardwoods (Wagner et al., 2012; Umezawa, 2010). The biosynthesis of monolignols in conifers starts with deamination of L-phenylalanine by phenylalanine ammonia lyase (PAL; EC 4.3.1.5) producing *p*-coumarate, which is the common precursor for the lignin and flavonoids biosynthetic pathways (Boerjan et al., 2003; Bagal et al., 2011). Most enzymes involved in monolignol biosynthesis in conifers seem to be encoded by multi-gene families, and recently multiple *PAL* genes were discovered in the LP genome (Bagal et al., 2011; Whetten and Sederoff, 1992). During our analysis, we identified a *PAL* gene which is 2.1 kb in length and shares 98% sequence identity to LP, and 97% sequence identity to *Pinus massoniana*, *Pinus pinaster*, and *Pinus tabuliformis*. Unlike angiosperm *PAL* genes, which include one intron and two exons, we observed no introns in the LP *PAL* gene (single exonic gene) (Fig. 2), which is consistent with the previous studies where *PAL* genes from *Pinus banksiana* (Bagal et al., 2011) and *Picea glauca* (Stival Sena et al., 2014) were shown to be intronless. Also, intronless *PAL* genes were observed in *Ginkgo biloba*, which implies intronless *PAL* genes might be a unique feature in gymnosperms (Xu et al., 2008). However, an intronless *PAL* gene in *Bambusa oldhamii* was discovered recently (first intronless *PAL* gene reported in angiosperms) (Hsieh et al., 2011). Since introns may have a significant effect on expression profile of a particular gene, such as enhanced transcriptional efficiency, the significance of presence vs. absence of intron(s) in this gene is of interest (Le Hir et al., 2003).

(F) *Caffeoyl-CoA 3-O-methyltransferase* (*CCoAOMT*) – One of the key enzymes involved in the biosynthesis of monolignols is caffeoyl-CoA 3-*O*-methyltransferase (CCoAOMT, EC 2.1. 1.104) which methylates caffeoyl-CoA (the preferred substrate for CCoAOMT) to produce feruloyl-CoA, which is primarily involved in biosynthesis of G-type lignin in coniferous gymnosperms-(Wagner et al., 2011). Suppression of *CCoAOMT* expression in angiosperm species caused a 20–45% reduction in lignin content (mainly reduction of G- and S-lignin) indicating that CCoAOMT is the precursor for both lignin types in angiosperms (Wagner et al., 2012). In contrast, in CCoAOMT-suppressed *Pinus radiata* transgenic lines, a 20% reduction of lignin content was observed with marked decrease in G-type lignin and increase in H type lignin, resulting in a 10-fold increase in the H/G ratio (Wagner et al., 2011). During our analysis, we discovered a *CCoAOMT* gene which is 3 kb in length and shares 94% sequence identity to LP and *Pinus pinaster* EST sequences. The genomic sequence of the LP *CCoAOMT* gene was composed of five exons separated by four introns (Fig. 2), and this structure is consistent with most of the angiosperms, which have from two to four exons (Barakat et al., 2011; Guillet-Claude et al., 2004).

(G) *Laccase* (*lac*) – During lignification, monolignol precursors (i.e., H, G, and, S units of lignin) synthesized in the cytoplasm are transported to the cell wall, where they are oxidized by laccases and/or peroxidases to initiate the polymerization process (Berthet et al., 2012). Laccases, which are members of multicopper oxidase family, are among the most abundant proteins in lignin-rich *Pinus radiata* compression wood, suggesting the importance of laccases in conifer lignin polymerization (Mast et al., 2010). Purified laccase from cell walls of LP has been shown to be involved in oxidation of coniferyl alcohol in vitro (Bao et al., 1993; O'Malley et al., 1993). Nevertheless, determining the specific role(s) of laccases/peroxidases in lignin polymerization process is far from complete. We identified an apparent *lac8* gene, which is 3.4 kb in length and shares 99% and 75% sequence identity to two LP *lac* genes (i.e., *lac8* and *lac7*, respectively). The *lac8* gene from our BAC is structured with six exons and five introns (Fig. 2), and shows structural similarity to angiosperm laccase homologs that are composed of four to six exons (Liang et al., 2006; McCaig et al., 2005; Gavnholt et al., 2002; Tomkova et al., 2012).

(H) *MYB transcription factor* (*MYB*) – The regulation of lignin biosynthesis is complex and involves a large number of transcriptional regulators. The vast majority of these transcriptional activators and repressors belong to the MYB family, one of the largest families of plant transcription factors (Zhong and Ye, 2009; Zhao and Dixon, 2011). The identified MYB transcription factors likely activate

expression of lignin-related genes in conifers by binding to AC elements present in the promoter region of lignin-related genes which are expressed in developing wood that undergoes secondary cell wall thickening and lignin biosynthesis (Bomal et al., 2008; Patzlaff et al., 2003a; Patzlaff et al., 2003b; Bedon et al., 2007). For example, LP *PtMYB4* (which is homologous to *Arabidopsis AtMYB46*) induced the expression of certain lignin biosynthetic genes leading to ectopic lignin deposition and increased secondary cell wall thickening in transgenic tobacco (*Nicotiana tabacum*) (Patzlaff et al., 2003a). Similar phenotypic effects were observed in transgenic spruce (*Picea glauca*) which overexpressed LP *PtMYB1* and *PtMYB8* (which are the respective homologs of *Arabidopsis AtMYB85* and *AtMYB46*) (Zhao and Dixon, 2011; Bomal et al., 2008). During our analysis, we sequenced the *MYB8* gene which is 2.4 kb in length and shares 99% sequence identity to LP and 98% sequence identity to *Pinus pinaster* ESTs. The genomic sequence of *MYB8* comprises three exons separated by two introns (Fig. 2), which is consistent with previous studies of angiosperms and some other gymnosperms, such as *Picea glauca*, where variable numbers of introns (i.e. zero to three) were observed (Stival Sena et al., 2014; Bedon et al., 2007).

3.9. Intron and exons

For the ten genes described above, the average number of exons per gene, average lengths of exons and introns, and percentages of exonic and intronic regions were determined. The average number of exons per gene (7.3) was smaller than the average number of exons (9.25) of some secondary cell-wall formation and nitrogen metabolism genes in *Picea glauca* and some angiosperm species (Stival Sena et al., 2014). For the majority of the large genes that we characterized (i.e. *CCoAOMT*, *CesA1*, *CesA2*, and *SuSy1*), percentages of intronic sequences are larger than the percentages of exonic sequences. Additionally, presence of large amounts of intronic sequence per gene in *Picea glauca* has been reported (Stival Sena et al., 2014). This has largely been due to the presence of long introns and repetitive element sequences which are ubiquitously harbored in introns (Stival Sena et al., 2014; Nystedt et al., 2013). Accumulation of one or a few longer introns, a conserved feature, in species with larger genomes such as *Picea abies*, *Picea glauca*, and LP has been observed across gymnosperm taxa (Nystedt et al., 2013; Stival Sena et al., 2014; Wegrzyn et al., 2014). The longest reported intron length in this set of ten genes is 1.5 kb. Longer introns such as 159 kb and 68 kb have been previously reported in LP and *Picea glauca*, respectively (Wegrzyn et al., 2014; Nystedt et al., 2013). However, of the above ten genes characterized, median intron length was 159 bp which is similar to *Picea glauca* (155 bp) and other plant species (100–200 bp). While it is likely that long introns represent a relatively small fraction of overall intronic content in conifers (Stival Sena et al., 2014), the limited size of individual LP BAC clone inserts (ca. 100 kb) (Magbanua et al., 2011) would greatly limit the probability of detecting very long introns (i.e., a 159 kb intron will not be discovered by sequencing a 100 kb BAC). The significance of longer introns in the LP genome has yet to be understood, even though it is observed that enhanced gene expression is associated with increased intron length in rice (*Oryza sativa*) and *Arabidopsis thaliana* (Ren et al., 2006).

4. Conclusions

Cot analysis provided an overview of the LP genome. Of note, the genome is roughly 80% repetitive DNA. Single/low-copy (SL) DNA accounts for 10% of the genome; however, the majority of SL DNA is apparently gene/repeat-poor DNA. Analysis of sequence from 99 BAC clones using combinations of similarity-based and ab initio gene prediction approaches resulted in identification of 154 gene models of which ten had sufficient support to be declared genes. The ten verified genes, which were all targeted, are involved in LP wood formation and carbon metabolism. Of the ten genes putatively associated with wood-formation, four (*CesA1*, *CesA2*, *KOR,* and *SuSy1*) appear involved in cellulose biosynthesis, two (*GMP1* and *GMP2*) in hemicellulose biosynthesis, and four (*PAL*, *CCoAOMPT*, *Lac 8*, and *MYB 8*) have been associated with lignification pathways. The identification and characterization of these genes related to wood formation should provide important tools for the genetic manipulation of wood property traits in LP and other conifers.

Authors' contributions

DP annotated the BAC sequences and wrote the first draft of the manuscript. ZVM performed BAC library screening. ST and DM performed the Cot analysis. MA and PC performed computational data analysis. CJN selected the cDNAs to be used in macroarray screening. JS and JG performed BAC sequencing and assembly. DGP and JFDD developed the project and oversaw its progression. All authors (led by DGP) contributed to editing of the manuscript.

Acknowledgements

The authors thank Benjamin D. Bartlett and Seval Ozkan for preparation of macroarrays for BAC screening, and Dr. Michael Murray for the recipe for ASE buffer. This work was funded, in part, by National Science Foundation award DBI 0421717 to DGP. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Competing interests

The authors declare that they have no competing interests.

Appendix A. Additional Files

Additional Files can be found online at https://doi.org/10.1016/j.gene.2018.04.024.

References

Ahuja, M.R., Neale, D.B., 2005. Evolution of genome size in conifers. Silvae Genet. 54 (3), 126–137.
Amborella Genome Project, 2013. The *Amborella* genome and the evolution of flowering plants. Science 342 (6165), 1241089.
An, X., Chen, Z., Wang, J., Ye, M., Ji, L., Wang, J., Liao, W., Ma, H., 2014. Identification and characterization of the *Populus* sucrose synthase gene family. Gene 539 (1), 58–67.
Appenzeller, L., Doblin, M., Barreiro, R., Wang, H., Niu, X., Kollipara, K., Carrigan, L., Tomes, D., Chapman, M., Dhugga, K.S., 2004. Cellulose synthesis in maize: isolation and expression analysis of the cellulose synthase (CesA) gene family. Cellulose 11 (3–4), 287–299.
Bagal, U.R., Leebens-Mack, J.H., Lorenz, W.W., Dean, J.F.D., 2011. Phylogenomic analysis of the phenylalanine ammonia lyase gene family in loblolly pine (*Pinus taeda* L.). IEEE 69–74.
Badejo, A.A., Tanaka, N., Esaka, M., 2008. Analysis of GDP-D-mannose pyrophosphorylase gene promoter from acerola (*Malpighia glabra*) and increase in ascorbate content of transgenic tobacco expressing the acerola gene. Plant Cell Physiol. 49 (1), 126–132.
Bao, W., O'Malley, D.M., Whetten, R., Sederoff, R.R., 1993. A laccase associated with lignification in loblolly pine xylem. Science 260 (5108), 672–674.
Barakat, A., Choi, A., Yassin, N.B.M., Park, J.S., Sun, Z., Carlson, J.E., 2011. Comparative genomics and evolutionary analyses of the O-methyltransferase gene family in *Populus*. Gene 479 (1–2), 37–46.
Baroja-Fernández, E., Muñoz, F.J., Saikusa, T., Rodríguez-López, M., Akazawa, T., Pozueta-Romero, J., 2003. Sucrose synthase catalyzes the de novo production of ADPglucose linked to starch biosynthesis in heterotrophic tissues of plants. Plant Cell Physiol. 44 (5), 500–509.
Baud, S., Vaultier, M.N., Rochat, C., 2004. Structure and expression profile of the sucrose synthase multigene family in *Arabidopsis*. J. Exp. Bot. 55 (396), 397–409.
Bedon, F., Grima-Pettenati, J., Mackay, J., 2007. Conifer R2R3-MYB transcription factors: sequence analyses and gene expression in wood-forming tissues of white spruce (*Picea glauca*). BMC Plant Biol. 7, 17.
Bennett, M.D., Leitch, I.J., 2005. Nuclear DNA amounts in angiosperms: progress, problems and prospects. Ann. Bot. 95 (1), 45–90.

Berthet, S., Thevenin, J., Baratiny, D., Demont-Caulet, N., Debeaujon, I., Bidzinski, P., Leple, J.C., Huis, R., Hawkins, S., Gomez, L.D., Lapierre, C., Jouanin, L., 2012. Role of plant laccases in lignin polymerization. Adv. Bot. Res. 61, 145–172.
Bérubé, Y., Zhuang, J., Rungis, D., Ralph, S., Bohlmann, J., Ritland, K., 2007. Characterization of EST-SSRs in loblolly pine and spruce. Tree Genet. Genome 3, 251–259.
Birol, I., Raymond, A., Jackman, S.D., Pleasance, S., Coope, R., Taylor, G.A., Yuen, M.M., Keeling, C.I., Brand, D., Vandervalk, B.P., Kirk, H., Pandoh, P., Moore, R.A., Zhao, Y., Mungall, A.J., Jaquish, B., Yanchuk, A., Ritland, C., Boyle, B., Bousquet, J., Ritland, K., Mackay, J., Bohlmann, J., Jones, S.J., 2013. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. Bioinformatics 29 (12), 1492–1497.
Boerjan, W., Ralph, J., Baucher, M., 2003. Lignin biosynthesis. Annu. Rev. Plant Biol. 54, 519–546.
Bogunic, F., Muratovic, E., Brown, S.C., Siljak-Yakovlev, S., 2003. Genome size and base composition of five *Pinus* species from the Balkan region. Plant Cell Rep. 22 (1), 59–63.
Bomal, C., Bedon, F., Caron, S., Mansfield, S.D., Levasseur, C., Cooke, J.E.K., Blais, S., Tremblay, L., Morency, M.J., Pavy, N., Grima-Pettenati, J., Séguin, A., MacKay, J., 2008. Involvement of *Pinus taeda* MYB1 and MYB8 in phenylpropanoid metabolism and secondary cell wall biogenesis: a comparative in planta analysis. J. Exp. Bot. 59 (14), 3925–3939.
Britten, R.J., Graham, D.E., Neufeld, B.R., 1974. Analysis of repeating DNA sequences by reassociation. Methods Enzymol. 29 (0), 363–418.
Brown Jr., R.M., 1996. The biosynthesis of cellulose. J. Macromol. Sci. Pure Appl. Chem. 33 (10), 1345–1373.
Bunge, J., Chouvarine, P., Peterson, D.G., 2009. CotQuest: improved algorithm and software for nonlinear regression analysis of DNA reassociation kinetics data. Anal. Biochem. 388, 322–330.
Cairney, J., Zheng, L., Cowels, A., Hsiao, J., Zismann, V., Liu, J., Ouyang, S., Thibaud-Nissen, F., Hamilton, J., Childs, K., Pullman, G.S., Zhang, Y., Oh, T., Buell, C.R., 2006. Expressed sequence tags from loblolly pine embryos reveal similarities with angiosperm embryogenesis. Plant Mol. Biol. 62 (4–5), 485–501.
Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., Ware, D., Shiu, S.H., Childs, K.L., Sun, Y., Jiang, N., Yandell, M., 2014. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164 (2), 513–524.
Chen, A., He, S., Li, F., Li, Z., Ding, M., Liu, Q., Rong, J., 2012. Analyses of the sucrose synthase gene family in cotton: structure, phylogeny and expression patterns. BMC Plant Biol. 12.
Chourey, P.S., Taliercio, E.W., Carlson, S.J., Ruan, Y.L., 1998. Genetic evidence that the two isozymes of sucrose synthase present in developing maize endosperm are critical, one for cell wall integrity and the other for starch biosynthesis. Mol. Gen. Genet. 259 (1), 88–96.
Coleman, H.D., Ellis, D.D., Gilbert, M., Mansfield, S.D., 2006. Up-regulation of sucrose synthase and UDP-glucose pyrophosphorylase impacts plant growth and metabolism. Plant Biotechnol. J. 4 (1), 87–101.
Coleman, H.D., Yan, J., Mansfield, S.D., 2009. Sucrose synthase affects carbon partitioning to increase cellulose production and altered cell wall ultrastructure. Proc. Natl. Acad. Sci. U. S. A. 106 (31), 13118–13123.
Conklin, P.L., Norris, S.R., Wheeler, G.L., Williams, E.H., Smirnoff, N., Last, R.L., 1999. Genetic evidence for the role of GDP-mannose in plant ascorbic acid (vitamin C) biosynthesis. Proc. Natl. Acad. Sci. U. S. A. 96 (7), 4198–4203.
Cossu, R.M., Casola, C., Giacomello, S., Vidalis, A., Scofield, D.G., Zuccolo, A., 2017. LTR retrotransposons show low levels of unequal recombination and high rates of intraelement gene conversion in large plant genomes. Genome Biol. Evol. 9 (12), 3449–3462.
Dauwe, R., Robinson, A., Mansfield, S., 2011. Recent advances in proteomics and metabolomics in gymnosperms. In: Plomion, C., Bousquet, J., Kole, C.R. (Eds.), Genetics, Genomics and Breeding of Conifers. Science Publishers, Edenbridge (pp.).
Daystar, J., Gonzalez, R., Reeb, C., Venditti, R., Treasure, T., Abt, R., Kelley, S., 2014. Economics, environmental impacts, and supply chain analysis of cellulosic biomass for biofuels in the southern U.S.: pine, eucalyptus, unmanaged hardwoods, forest residues, switchgrass, and sweet sorghum. Bioresources 9 (1), 393–444.
Dhugga, K.S., Barreiro, R., Whitten, B., Stecca, K., Hazebroek, J., Randhawa, G.S., Dolan, M., Kinney, A.J., Tomes, D., Nichols, S., Anderson, P., 2004. Guar seed β-mannan synthase is a member of the cellulose synthase super gene family. Science 303 (5656), 363–366.
Djerbi, S., Lindskog, M., Arvestad, L., Sterky, F., Teeri, T.T., 2005. The genome sequence of black cottonwood (*Populus trichocarpa*) reveals 18 conserved cellulose synthase (CesA) genes. Planta 221 (5), 739–746.
El Baidouri, M., Panaud, O., 2013. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. Genome Biol. Evol. 5 (5), 954–965.
Endler, A., Persson, S., 2011. Cellulose synthases and synthesis in *Arabidopsis*. Mol. Plant 4 (2), 199–211.
Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8, 186–194.
Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8, 175–185.
Ferris, P., Olson, B.J.S.C., De Hoff, P.L., Douglass, S., Casero, D., Prochnik, S., Geng, S., Rai, R., Grimwood, J., Schmutz, J., Nishfi, I., Hamaji, T., Nozaki, H., Pellegrini, M., Umen, J.G., 2010. Evolution of an expanded sex-determining locus in *Volvox*. Science 328 (5976), 351–354.
Frederick Jr., W.J., Lien, S.J., Courchene, C.E., Demartini, N.A., Ragauskas, A.J., Iisa, K., 2008. Production of ethanol from carbohydrates from loblolly pine: a technical and economic assessment. Bioresour. Technol. 99, 5051–5057.
Fujii, S., Hayashi, T., Mizuno, K., 2010. Sucrose synthase is an integral component of the cellulose synthesis machinery. Plant Cell Physiol. 51 (2), 294–301.
Gavnholt, B., Larsen, K., Rasmussen, S.K., 2002. Isolation and characterisation of laccase cDNAs from meristematic and stem tissues of ryegrass (*Lolium perenne*). Plant Sci. 162 (6), 873–885.
Galbe, M., Zacchi, G., 2002. A review of the production of ethanol from softwood. Appl. Microbiol. Biotechnol. 59 (6), 618–628.
Geigenberger, P., Reimholz, R., Geiger, M., Merlo, L., Canale, V., Stitt, M., 1997. Regulation of sucrose and starch metabolism in potato tubers in response to short-term water deficit. Planta 201 (4), 502–518.
Gordon, D., Abajian, C., Green, P., 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8, 195–202.
Gordon, A.J., Minchin, F.R., James, C.L., Komina, O., 1999. Sucrose synthase in legume nodules is essential for nitrogen fixation. Plant Physiol. 120 (3), 867–877.
Guan, R., Zhao, Y., Zhang, H., Fan, G., Liu, X., Zhou, W., Shi, C., Wang, J., Liu, W., Liang, X., Fu, Y., Ma, K., Zhao, L., Zhang, F., Lu, Z., Lee, S.M., Xu, X., Wang, J., Yang, H., Fu, C., Ge, S., Chen, W., 2016. Draft genome of the living fossil *Ginkgo biloba*. GigaScience 5 (1), 49.
Guerriero, G., Fugelstad, J., Bulone, V., 2010. What do we really know about cellulose biosynthesis in higher plants? J. Integr. Plant Biol. 52 (2), 161–175.
Guillet-Claude, C., Birolleau-Touchard, C., Manicacci, D., Fourmann, M., Barraud, S., Carret, V., Martinant, J.P., Barrière, Y., 2004. Genetic diversity associated with variation in silage corn digestibility for three O-methyltransferase genes involved in lignin biosynthesis. Theor. Appl. Genet. 110 (1), 126–135.
Harris, D., DeBolt, S., 2010. Synthesis, regulation and utilization of lignocellulosic biomass. Plant Biotechnol. J. 8 (3), 244–262.
Hazen, S.P., Scott-Craig, J.S., Walton, J.D., 2002. Cellulose synthase-like genes of rice. Plant Physiol. 128 (2), 336–340.
Hirose, T., Scofield, G.N., Terao, T., 2008. An expression analysis profile for the entire sucrose synthase gene family in rice. Plant Sci. 174 (5), 534–543.
Hizume, M., Kondo, T., Shibata, F., Ishizuka, R., 2001. Flow cytometric determination of genome size in the Taxodiaceae, Cupressaceae sensu stricto and Sciadopityaceae. Cytologia 66, 307–311.
Holt, C., Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinforma. 12 (1).
Horst, I., Welham, T., Kelly, S., Kaneko, T., Sato, S., Tabata, S., Parniske, M., Wang, T.L., 2007. Tilling mutants of *Lotus japonicus* reveal that nitrogen assimilation and fixation can occur in the absence of nodule-enhanced sucrose synthase. Plant Physiol. 144 (2), 806–820.
Hsieh, L.S., Hsieh, Y.L., Yeh, C.S., Cheng, C.Y., Yang, C.C., Lee, P.D., 2011. Molecular characterization of a phenylalanine ammonia-lyase gene (BoPAL1) from *Bambusa oldhamii*. Mol. Biol. Rep. 38 (1), 283–290.
Joshi, C.P., Bhandari, S., Ranjan, P., Kalluri, U.C., Liang, X., Fujino, T., Samuga, A., 2004. Genomics of cellulose biosynthesis in poplars. New Phytol. 164 (1), 53–61.
Kirst, M., Johnson, A.F., Baucom, C., Ulrich, E., Hubbard, K., Staggs, R., Paule, C., Retzel, E., Whetten, R., Sederoff, R.R., 2003. Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U. S. A. 100, 7383–7388.
Kelly, L.J., Renny-Byfield, S., Pellicer, J., Macas, J., Novak, P., Neumann, P., Lysak, M.A., Day, P.D., Berger, M., Fay, M.F., Nichols, R.A., Leitch, A.R., Leitch, I.J., 2015. Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytol. 208 (2), 596–607.
Kovach, A., Wegrzyn, J.L., Parra, G., Holt, C., Bruening, G.E., Loopstra, C.A., Hartigan, J., Yandell, M., Langley, C.H., Korf, I., Neale, D.B., 2010. The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics 11, 420.
Le Hir, H., Nott, A., Moore, M.J., 2003. How introns influence and enhance eukaryotic gene expression. Trends Biochem. Sci. 28 (4), 215–220.
Lerouxel, O., Cavalier, D.M., Liepman, A.H., Keegstra, K., 2006. Biosynthesis of plant cell wall polysaccharides - a complex process. Curr. Opin. Plant Biol. 9 (6), 621–630.
Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H., Barker, M.S., 2015. Early genome duplications in conifers and other seed plants. Sci. Adv. 1 (10), e1501084.
Liang, M., Haroldsen, V., Cai, X., Wu, Y., 2006. Expression of a putative laccase gene, ZmLAC1, in maize primary roots under stress. Plant Cell Environ. 29 (5), 746–753.
Liepman, A.H., Nairn, C.J., Willats, W.G.T., Sørensen, I., Roberts, A.W., Keegstra, K., 2007. Functional genomic analysis supports conservation of function among cellulose synthase-like a gene family members and suggests diverse roles of mannans in plants. Plant Physiol. 143 (4), 1881–1893.
Liu, W., Thummasuwan, S., Sehgal, S.K., Chouvarine, P., Peterson, D.G., 2011. Characterization of the genome of bald cypress. BMC Genomics 12 (1), 553.
Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., Borodovsky, M., 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33 (20), 6494–6506.
Lorenz, W.W., Sun, F., Liang, C., Kolychev, D., Wang, H., Zhao, X., Cordonnier-Pratt, M.-M., Pratt, L.H., Dean, J.F., 2006. Water stress-responsive genes in loblolly pine (*Pinus taeda*) roots identified by analyses of expressed sequence tag libraries. Tree Physiol. 26, 1–16.
Lorenz, W.W., Ayyampalayam, S., Bordeaux, J.M., Howe, G.T., Jermstad, K.D., Neale, D.B., Rogers, D.L., Dean, J.F.D., 2012. Conifer DBMagic: a database housing multiple de novo transcriptome assemblies for 12 diverse conifer species. Tree Genet. Genome 8 (6), 1477–1485.
MacKay, J., Dean, J.F., Plomion, C., Peterson, D.G., Canovas, F.M., Pavy, N., Ingvarsson,

P.K., Savolainen, O., Guevara, M.A., Fluch, S., Vinceti, B., Abarca, D., Diaz-Sala, C., Cervera, M.T., 2012. Towards decoding the conifer giga-genome. Plant Mol. Biol. 80 (6), 555–569.
Magbanua, Z.V., Ozkan, S., Bartlett, B.D., Chouvarine, P., Saski, C.A., Liston, A., Cronn, R.C., Nelson, C.D., Peterson, D.G., 2011. Adventures in the enormous: a 1.8 million clone BAC library for the 21.7 Gb genome of loblolly pine. PLoS One 6 (1), e16214.
Mandel, M., Marmur, J., 1968. Use of ultraviolet absorbance-temperature profile for determining the guanine plus cytosine content of DNA. Methods Enzymol. 12, 195–206.
Mast, S., Peng, L., Jordan, T.W., Flint, H., Phillips, L., Donaldson, L., Strabala, T.J., Wagner, A., 2010. Proteomic analysis of membrane preparations from developing *Pinus radiata* compression wood. Tree Physiol. 30 (11), 1456–1468.
McCaig, B.C., Meagher, R.B., Dean, J.F.D., 2005. Gene structure and molecular analysis of the laccase-like multicopper oxidase (LMCO) gene family in *Arabidopsis thaliana*. Planta 221 (5), 619–636.
Mølhøj, M., Pagant, S., Höfte, H., 2002. Towards understanding the role of membrane-bound endo-β-1,4-glucanases in cellulose biosynthesis. Plant Cell Physiol. 43 (12), 1399–1406.
Maloney, V.J., Mansfield, S.D., 2010. Characterization and varied expression of a membrane-bound endo-β-1,4-glucanase in hybrid poplar. Plant Biotechnol. J. 8 (3), 294–307.
Maloney, V.J., Samuels, A.L., Mansfield, S.D., 2012. The endo-1,4-β-glucanase Korrigan exhibits functional conservation between gymnosperms and angiosperms and is required for proper cell wall formation in gymnosperms. New Phytol. 193 (4), 1076–1087.
Morse, A.M., Peterson, D.G., Islam-Faridi, M.N., Smith, K.E., Magbanua, Z., Garcia, S.A., Kubisiak, T.L., Amerson, H.V., Carlson, J.E., Nelson, C.D., Davis, J.M., 2009. Evolution of genome size and complexity in *Pinus*. PLoS One 4 (2), e4332.
Murray, B.G., Leitch, I.J., Bennett, M.D., 2012. Gymnosperm DNA C-values Database (Release 5.0, Dec. 2012) Accessed on Jan. 28, 2017. http://www.kew.org/cvalues/.
Nairn, C.J., Haselkorn, T., 2005. Three loblolly pine CesA genes expressed in developing xylem are orthologous to secondary cell wall CesA genes of angiosperms. New Phytol. 166 (3), 907–915.
Nairn, C.J., Lennon, D.M., Wood-Jones, A., Nairn, A.V., Dean, J.F.D., 2008. Carbohydrate-related genes and cell wall biosynthesis in vascular tissues of loblolly pine (*Pinus taeda*). Tree Physiol. 28 (7), 1099–1110.
Neale, D.B., Wegrzyn, J.L., Stevens, K.A., Zimin, A.V., Puiu, D., Crepeau, M.W., Cardeno, C., Koriabine, M., Holtz-Morris, A.E., Liechty, J.D., Martínez-García, P.J., Vasquez-Gross, H.A., Lin, B.Y., Zieve, J.J., Dougherty, W.M., Fuentes-Soriano, S., Wu, L.S., Gilbert, D., Marçais, G., Roberts, M., Holt, C., Yandell, M., Davis, J.M., Smith, K.E., Dean, J.F.D., Lorenz, W.W., Whetten, R.W., Sederoff, R., Wheeler, N., McGuire, P.E., Main, D., Loopstra, C.A., Mockaitis, K., deJong, P.J., Yorke, J.A., Salzberg, S.L., Langley, C.H., 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 15 (3), R59.
Neves, L.G., Davis, J.M., Barbazuk, W.B., Kirst, M., 2014. A high-density gene map of loblolly pine (*Pinus taeda* L.) based on exome sequence capture genotyping. G3 Genes Genome Genet 4 (1), 29–37.
Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., Koriabine, M., Kucukoglu, M., Käller, M., Luthman, J., Lysholm, F., Niittylä, T., Olson, A., Rilakovic, N., Ritland, C., Rosselló, J.A., Sena, J., Svensson, T., Talavera-López, C., Theißen, G., Tuominen, H., Vanneste, K., Wu, Z.Q., Zhang, B., Zerbe, P., Arvestad, L., Bhalerao, R., Bohlmann, J., Bousquet, J., Garcia Gil, R., Hvidsten, T.R., De Jong, P., MacKay, J., Morgante, M., Ritland, K., Sundberg, B., Thompson, S.L., Van De Peer, Y., Andersson, B., Nilsson, O., Ingvarsson, P.K., Lundeberg, J., Jansson, S., 2013. The Norway spruce genome sequence and conifer genome evolution. Nature 497 (7451), 579–584.
O'Brien, I.E.W., Smith, D.R., Gardner, R.C., Murray, B.G., 1996. Flow cytometric determination of genome size in *Pinus*. Plant Sci. 115, 91–99.
O'Malley, D.M., Whetten, R., Bao, W., Chen, C.L., Sederoff, R.R., 1993. The role of laccase in lignification. Plant J. 4 (5), 751–757.
Palle, S.R., Seeve, C.M., Eckert, A.J., Cumbie, W.P., Goldfarb, B., Loopstra, C.A., 2011. Natural variation in expression of genes involved in xylem development in loblolly pine (*Pinus taeda* L.). Tree Genet. Genome 7 (1), 193–206.
Patzlaff, A., McInnis, S., Courtenay, A., Surman, C., Newman, L.J., Smith, C., Bevan, M.W., Mansfield, S., Whetten, R.W., Sederoff, R.R., Campbell, M.M., 2003a. Characterisation of a pine MYB that regulates lignification. Plant J. 36 (6), 743–754.
Patzlaff, A., Newman, L.J., Dubos, C., Whetten, R.W., Smith, C., McInnis, S., Bevan, M.W., Sederoff, R.R., Campbell, M.M., 2003b. Characterisation of PtMYB1, an R2R3-MYB from pine xylem. Plant Mol. Biol. 53 (4), 597–608.
Pauly, M., Keegstra, K., 2008. Cell-wall carbohydrates and their modification as a resource for biofuels. Plant J. 54 (4), 559–568.
Pavy, N., Pelgas, B., Beauseigle, S., Blais, S., Gagnon, F., Gosselin, I., Lamothe, M., Isabel, N., Bousquet, J., 2008. Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics 9, 21.
Pear, J.R., Kawagoe, Y., Schreckengost, W.E., Delmer, D.P., Stalker, D.M., 1996. Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. Proc. Natl. Acad. Sci. U. S. A. 93 (22), 12637–12642.
Pearson, W.R., Davidson, E.H., Britten, R.J., 1977. A program for least squares analysis of reassociation and hybridization data. Nucleic Acids Res. 4, 1727–1737.
Peterson, D.G., 2005. Reduced representation strategies and their application to plant genomes. In: Meksem, K., Kahl, G. (Eds.), The Handbook of Genome Mapping: Genetic and Physical Mapping. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, pp. 307–335.
Peterson, D.G., Boehm, K.S., Stack, S., 1997. Isolation of milligram quantities of DNA from tomato (*Lycopersicon esculentum*), a plant containing high levels of polyphenolic compounds. Plant Mol. Biol. Report. 15, 148–153.
Peterson, D.G., Pearson, W.R., Stack, S.M., 1998. Characterization of the tomato (*Lycopersicon esculentum*) genome using in vitro and in situ DNA reassociation. Genome 41, 346–356.
Peterson, D.G., Schulze, S.R., Sciara, E.B., Lee, S.A., Bowers, J.E., Nagel, A., Jiang, N., Tibbitts, D.C., Wessler, S.R., Paterson, A.H., 2002. Integration of cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res. 12 (5), 795–807.
Pellicer, J., Hidalgo, O., Dodsworth, S., Leitch, I.J., 2018. Genome size diversity and its impact on the evolution of land plants. Genes (Basel) 9 (2), 88.
Plomion, C., Chagné, D., Pot, D., Kumar, S., Wilcox, P.L., Burdon, R.D., Prat, D., Peterson, D.G., Paiva, J., Chaumeil, P., Vendramin, G.G., Sebastiani, F., Nelson, C.D., Echt, C.S., Savolainen, O., Kubisiak, T.L., Cervera, M.T., de María, N., Islam-Faridi, M.N., 2007. The pines. In: Kole, C.R. (Ed.), Genome Mapping and Molecular Breeding in Plants, Vol. 7 - Forest Trees. Springer, Heidelberg, Berlin, New York, Tokyo, pp. 29–78.
Rake, A.V., Miksche, J.P., Hall, R.B., Hansen, K.M., 1980. DNA reassociation kinetics of four conifers. Can. J. Genet. Cytol. 22, 69–79.
Ranik, M., Myburg, A.A., 2006. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. Tree Physiol. 26 (5), 545–556.
Ren, X.Y., Vorst, O., Fiers, M.W.E.J., Stiekema, W.J., Nap, J.P., 2006. In plants, highly expressed genes are the least compact. Trends Genet. 22 (10), 528–532.
Ritland, K., 2012. Genomics of a phylum distant from flowering plants: conifers. Tree Genet. Genome 8 (3), 573–582.
Ralph, J., MacKay, J.J., Hatfield, R.D., O'Malley, D.M., Whetten, R.W., Sederoff, R.R., 1997. Abnormal lignin in a loblolly pine mutant. Science 277 (5323), 235–239.
Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., Barrell, B., 2000. Artemis: sequence visualization and annotation. Bioinformatics 16 (10), 944–945.
Saxena, I.M., Brown Jr., R.M., 2005. Cellulose biosynthesis: current views and evolving concepts. Ann. Bot. 96 (1), 9–21.
Scheller, H.V., Ulvskov, P., 2010. Hemicelluloses. Annu. Rev. Plant Biol. 61, 263–289.
Schubert, I., Vu, G.T., 2016. Genome stability and evolution: attempting a holistic view. Trends Plant Sci. 21 (9), 749–757.
Seoane-Zonjic, P., Cañas, R.A., Bautista, R., Gómez-Maldonado, J., Arrillaga, I., Fernández-Pozo, N., Claros, M.G., Cánovas, F.M., Ávila, C., 2016. Establishing gene models from the *Pinus pinaster* genome using gene capture and BAC sequencing. BMC Genomics 17 (1).
Smith, C.D., Edgar, R.C., Yandell, M.D., Smith, D.R., Celniker, S.E., Myers, E.W., Karpen, G.H., 2007. Improved repeat identification and masking in Dipterans. Gene 389 (1), 1–9.
Solovyev, V., Kosarev, P., Seledsov, I., Vorobyev, D., 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. 7 (Suppl. 1), S10.11–S10.12.
Slater, G.S., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinforma. 6, 31.
Smit, A.F.A., Hubley, R., Green, P., 2013. RepeatMasker Open-4.0. 2013-2015. http://www.repeatmasker.org, Accessed date: 15 March 2018.
Stainback, G.A., Alavalapati, J.R.R., 2002. Economic analysis of slash pine forest carbon sequestration in the southern U.S. J. For. Econ. 8 (2), 105–117.
Stanke, M., Steinkamp, R., Waack, S., Morgenstern, B., 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32, W309–W312.
Stival Sena, J., Giguère, I., Boyle, B., Rigault, P., Birol, I., Zuccolo, A., Ritland, K., Ritland, C., Bohlmann, J., Jones, S., Bousquet, J., Mackay, J., 2014. Evolution of gene structure in the conifer *Picea glauca*: a comparative analysis of the impact of intron size. BMC Plant Biol. 14 (1).
Suzuki, S., Li, L., Sun, Y.H., Chiang, V.L., 2006. The cellulose synthase gene superfamily and biochemical functions of xylem-specific cellulose synthase-like genes in *Populus trichocarpa*. Plant Physiol. 142 (3), 1233–1245.
Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., Huala, E., 2008. The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res. 36 (Database issue), D1009–D1014.
Szyjanowicz, P.M.J., McKinnon, I., Taylor, N.G., Gardiner, J., Jarvis, M.C., Turner, S.R., 2004. The irregular xylem 2 mutant is an allele of Korrigan that affects the secondary cell wall of *Arabidopsis thaliana*. Plant J. 37 (5), 730–740.
Tarailo-Graovac, M., Chen, N., 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinforma. Suppl. 25, 4.10.11–14.10.14.
Tomkova, L., Kucera, L., Vaculova, K., Milotova, J., 2012. Characterization and mapping of a putative laccase-like multicopper oxidase gene in the barley (*Hordeum vulgare* L.). Plant Sci. 183, 77–85.
Umezawa, T., 2010. The cinnamate/monolignol pathway. Phytochem. Rev. 9 (1), 1–17.
Wagner, A., Tobimatsu, Y., Phillips, L., Flint, H., Torr, K., Donaldson, L., Pears, L., Ralph, J., 2011. CCoAOMT suppression modifies lignin composition in *Pinus radiata*. Plant J. 67 (1), 119–129.
Wagner, A., Donaldson, L., Ralph, J., 2012. Lignification and lignin manipulations in conifers. Adv. Bot. Res. 61, 37–76.
Wan, T., Liu, Z.M., Li, L.F., Leitch, A.R., Leitch, I.J., Lohaus, R., Liu, Z.J., Xin, H.P., Gong, Y.B., Liu, Y., Wang, W.C., Chen, L.Y., Yang, Y., Kelly, L.J., Yang, J., Huang, J.L., Li, Z., Liu, P., Zhang, L., Liu, H.M., Wang, H., Deng, S.H., Liu, M., Li, J., Ma, L., Liu, Y., Lei, Y., Xu, W., Wu, L.Q., Liu, F., Ma, Q., Yu, X.R., Jiang, Z., Zhang, G.Q., Li, S.H., Li, R.Q., Zhang, S.Z., Wang, Q.F., Van de Peer, Y., Zhang, J.B., Wang, X.M., 2018. A genome for gnetophytes and early evolution of seed plants. Nat. Plants 4 (2), 82–89.

Wang, L., Guo, K., Li, Y., Tu, Y., Hu, H., Wang, B., Cui, X., Peng, L., 2010. Expression profiling and integrative analysis of the CESA/CSL superfamily in rice. BMC Plant Biol. 10, 282.
Wegrzyn, J.L., Lin, B.Y., Zieve, J.J., Dougherty, W.M., Martínez-García, P.J., Koriabine, M., Holtz-Morris, A., deJong, P., Crepeau, M., Langley, C.H., Puiu, D., Salzberg, S.L., Neale, D.B., Stevens, K.A., 2013. Insights into the loblolly pine genome: characterization of BAC and fosmid sequences. PLoS One 8 (9).
Wegrzyn, J.L., Liechty, J.D., Stevens, K.A., Wu, L.S., Loopstra, C.A., Vasquez-Gross, H.A., Dougherty, W.M., Lin, B.Y., Zieve, J.J., Martínez-García, P.J., Holt, C., Yandell, M., Zimin, A.V., Yorke, J.A., Crepeau, M.W., Puiu, D., Salzberg, S.L., de Jong, P.J., Mockaitis, K., Main, D., Langley, C.H., Neale, D.B., 2014. Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. Genetics 196 (3), 891–909.
Whetten, R.W., Sederoff, R.R., 1992. Phenylalanine ammonia-lyase from loblolly pine: purification of the enzyme and isolation of complementary DNA clones. Plant Physiol. 98 (1), 380–386.
Whetten, R., Sun, Y.H., Zhang, Y., Sederoff, R., 2001. Functional genomics and cell wall biosynthesis in loblolly pine. Plant Mol. Biol. 47 (1–2), 275–291.
Xu, F., Cai, R., Cheng, S., Du, H., Wang, Y., Cheng, S., 2008. Molecular cloning, characterization and expression of phenylalanine ammonia-lyase gene from *Ginkgo biloba*. Afr. J. Biotechnol. 7 (6), 721–729.
Xu, S.M., Brill, E., Llewellyn, D.J., Furbank, R.T., Ruan, Y.L., 2012. Overexpression of a potato sucrose synthase gene in cotton accelerates leaf expansion, reduces seed abortion, and enhances fiber production. Mol. Plant 5 (2), 430–441.
Yang, S.H., Van Zyl, L., No, E.G., Loopstra, C.A., 2004. Microarray analysis of genes preferentially expressed in differentiating xylem of loblolly pine (*Pinus taeda*). Plant Sci. 166 (5), 1185–1195.
Yandell, M., Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. 13 (5), 329–342.
Yu, L., Chen, H., Sun, J., Li, L., 2014. PtrKOR1 is required for secondary cell wall cellulose biosynthesis in *Populus*. Tree Physiol. 34 (11), 1289–1300.
Zhao, Q., Dixon, R.A., 2011. Transcriptional networks for lignin biosynthesis: more complex than we thought? Trends Plant Sci. 16 (4), 227–233.
Zhong, R., Ye, Z.H., 2009. Transcriptional regulation of lignin biosynthesis. Plant Signal. Behav. 4 (11), 1028–1034.
Zimin, A., Stevens, K.A., Crepeau, M.W., Holtz-Morris, A., Koriabine, M., Marçais, G., Puiu, D., Roberts, M., Wegrzyn, J.L., de Jong, P.J., Neale, D.B., Salzberg, S.L., Yorke, J.A., Langley, C.H., 2014. Sequencing and assembly of the 22-Gb loblolly pine genome. Genetics 196 (3), 875–890.
Zimin, A.V., Stevens, K.A., Crepeau, M.W., Puiu, D., Wegrzyn, J.L., Yorke, J.A., Langley, C.H., Neale, D.B., Salzberg, S.L., 2017. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. GigaScience 6 (1), giw016.
Zuo, J., Niu, Q.W., Nishizawa, N., Wu, Y., Kost, B., Chua, N.H., 2000. KORRIGAN, an *Arabidopsis* endo-1,4-β-glucanase, localizes to the cell plate by polarized targeting and is essential for cytokinesis. Plant Cell 12 (7), 1137–1152.